Network: Computation in Neural Systems 5 (1994) 517–548. Printed in the UK

The statistics of natural images

Daniel L Ruderman†

Physiological Laboratory, University of Cambridge, Downing Street, Cambridge CB2 3EG, UK

Received 26 July 1994

Abstract. Recently there has been a resurgence of interest in the properties of natural images. Their statistics are important not only in image compression but also for the study of sensory processing in biology, which can be viewed as satisfying certain 'design criteria'. This review summarizes previous work on image statistics and presents our own data. Perhaps the most notable property of natural images is an invariance to scale. We present data to support this claim as well as evidence for a hierarchical invariance in natural scenes. These symmetries provide a powerful description of natural images as they greatly restrict the class of allowed distributions.

1. Introduction

We can easily distinguish images of the natural world from man-made pictures or those created randomly by a computer. Natural images are distinctive because they contain particular types of structure. They are far from random: images constructed randomly on a computer practically never contain a naturalistic scene—or even a tree. Natural images are thus very rare among the huge space of all possible images. How can we make use of this fact?

Image processing systems such as compression algorithms, analogue storage media, and our own visual system work with these real-world images. Thus to understand the typical behaviour of these systems we must first study the structure of natural scenes. We can then address questions like 'How much compression should we expect to achieve?' or 'How often will the playback distortion be above a critical value?' or 'How many bits of information per second does the optic nerve deliver?'

In this review we begin by summarizing what is known about the characteristics of natural images and add new findings. The study is motivated by practical examples of the use of such image statistics, primarily toward biological vision. In analysing images we ask the following questions:

- In which ways do natural images differ from random images?
- Do natural image statistics obey any simple invariances?
- What implications do these statistics have for image processing in biological visual systems?

In the first section we present an overview of statistical methods in signal processing, with an emphasis on applications to vision. A framework is described for understanding visual performance in terms of design criteria which involve the statistics of natural scenes.

† E-mail: dlr1002@cus.cam.ac.uk

0954-898X/94/040517+32$19.50

Subsequently, we outline a few ways in which these images can be characterized, and present previous work which hints at some simple properties underlying the structure of natural images. Finally, we perform a detailed analysis of our data and try to quantify the amount of predictability or redundancy present in the images. We are able to confirm the scaling which was suggest by other work, and discover a new invariance related to the hierarchical structure of natural scenes.

2. The statistical framework

The images we encounter every day comprise a very sparse subset of all possible images. Most images simply never appear in nature, as can easily be demonstrated by creating random images on a computer. Imagine all 256×256 images as existing in a 65536-dimensional space, then the 'volume' of the space which is occupied by natural images is infinitesimally small. Furthermore, natural images are not Gaussian, nor are they drawn from any other elementary distribution of the type commonly used in image modelling. In Field's words [32]

> '...the state-space describing the probability density of natural scenes is highly predictable but does not have the shape that is widely presumed.'

The distribution of natural images is complicated. Perhaps it is something like beer foam, which is mostly empty but contains a thin meshwork of fluid which fills the space and occupies almost no volume. The fluid region represents those images which are natural in character. Our intuition should be that there is no 'simple' transformation of the space which removes the distribution's complexity. The fact that the space is largely unoccupied expresses the redundancy in the distribution.

Since natural images are highly non-random, we might suspect that expressing them on a pixel-for-pixel basis is not the most convenient choice, as it is for entirely random images. If there is a way to encode the most frequently occuring images as short strings of bits and the least likely images as longer strings (perhaps using a Huffman code [41]), then the amount of storage space required *on average* can be reduced. Most practical algorithms don't work at the whole image level, but instead consider subimages such as horizontal scan lines (predictive coding) or 8×8 pixel blocks (JPEG). These are then encoded as independent entities, ignoring their interdependencies. The fact that such procedures are currently saving an order of magnitude in disk space [1] speaks for the large amount of redundancy and predictability contained in real images. Most importantly, it is the statistical structure of these images which determines the best compression algorithm. But at this stage we do not even know how much compression could be achieved in principle, since the statistics of real images have not as yet been well characterized.

Central to the discussion is the concept of an image *ensemble*. We imagine that each image $I(x)$ has associated with it a probability of occurence†, $P[I(x)]$, which defines the ensemble. Images drawn randomly from this distribution will completely represent the natural environment at hand. Practical questions might relate to the performance of a device or algorithm when acting on images drawn from this ensemble.

To give an illustrative example of the statistical formulation, we ask how well a set of images is represented by a noisy linear encoding. Let us simplify by considering a one-dimensional random signal $\phi(x)$ (band-limited and wide-sense stationary with zero

† Strictly speaking, we should invoke a probability density over the space of images (with an associated measure), since they comprise a continuum of 'events'.

ensemble mean) which is convolved with a filter $f(x)$ at each point. Random noise given by $\eta(x)$ is added to this signal. The final encoding, $y(x)$, may represent the signal recorded on analogue tape or the responses of an array of neurons. Specifically, we have

$$y(x) = (f * \phi)(x) + \eta(x) \tag{1}$$

and we wish to know how well $\phi(x)$ is represented by $y(x)$. One way to find out is by reconstructing the 'best' estimate of $\phi(x)$ from the signal $y(x)$†.

We define the best estimate, $\phi_{\text{est}}(x)$, to be the one with the minimum mean-squared error, that is

$$\langle |\phi_{\text{est}}(x) - \phi(x)|^2 \rangle \tag{2}$$

is minimized, where the expectation value is over all signals and noise. It is well known that this estimator is given by the mean of the posterior distribution[65], $P[\phi|y]$:

$$\phi_{\text{est}}(x) = \int \mathcal{D}\phi \; \phi(x) \; P[\phi|y] \tag{3}$$

where the integral is over all $\phi(x)$ in the ensemble.

The statistics of the ensemble enter via Bayes' theorem as a prior distribution:

$$P[\phi|y] = \frac{P[y|\phi]P[\phi]}{P[y]} \tag{4}$$

and $P[y|\phi]$ depends on the filter $f(x)$ and the noise statistics. So we have

$$\phi_{\text{est}}(x) = \frac{\int \mathcal{D}\phi \; \phi(x) P[y|\phi]P[\phi]}{\int \mathcal{D}\phi \; P[y|\phi]P[\phi]}. \tag{5}$$

This equation tells us how to go from the measurements $y(x)$ to a best guess of the signal, $\phi_{\text{est}}(x)$. When both the signal and the noise probability distributions are Gaussian the estimate is a linear functional of the measurements. But this is not true in general, and the estimator can be quite complicated. In the above formulation it is clear that knowledge of the ensemble, given by $P[\phi]$, is important to understanding how well the system will operate.

In general $P[\phi]$ must be fully characterized in order to do the above calculations. But this is impossible when faced with a high-dimensional signal as it would require gathering a huge number of images from the distribution‡. Fortunately there is a regime in which only a few of the correlation functions of the distribution are needed—when the signal-to-noise ratio (SNR) is low.

It can be shown that to lowest order in the SNR the best estimate is given by:

$$\Phi_{\text{est}}(k) = \frac{F(k)\mathcal{S}(k)}{\mathcal{N}(k)} Y(k) \tag{6}$$

where Φ_{est}, F, and Y are the Fourier transforms of ϕ_{est}, f, and y, respectively, and $\mathcal{S}(k)$ and $\mathcal{N}(k)$ are the ensemble power spectral densities of the signal and noise, respectively, at spatial frequency k. The power spectrum of the signal ensemble can be expressed in terms of its second-order correlation function as

$$\langle \Phi^*(k)\Phi(k') \rangle = 2\pi \mathcal{S}(k)\delta(k - k'). \tag{7}$$

† An overview of linear signal analysis and estimation can be found in the pioneering work of Wiener [88].

‡ Just imagine how many 4×4 images would need to be seen in order to fill out their probability distribution. Suppose each pixel is quantized to 16 grey values. Then there are over 10^{19} possible images, and we would need many times this many examples in order to make a guess as to how they are distributed.

At higher SNR equation (6) will include higher-order correlation functions of the image ensemble. Note that this equation is also the first-order term in a low SNR expansion of the Wiener filter.

We could now ask a question like: 'Which choice of $f(x)$ (within specified constraints) minimizes the expected reconstruction error?' Once the statistical character of the problem is specified, we are ready not only to quantify its typical behaviour, but also to ask what the system's optimal design is with respect to a given criterion. This paradigm has become popular in recent years, and has given rise to encouraging results when applied to biological vision, such as predicting neural responses [48, 50, 68], receptive fields [2, 4, 5, 10, 54, 57, 71, 72, 80, 83, 84, 86], colour coding [3, 8, 14, 22, 59, 64], stereo coding [53], the design of compound eyes [47, 78, 79], and even the pupil response [49].

The ultimate goal of the approach is to predict from first principles a measured response of a biological system, be it a neuron's activity, the optimum facet spacing in a compound eye, or a human psychophysical threshold. In the latter category, Atick and Redlich [4] have predicted the optimum neural encoding of natural scenes based on the criterion of minimizing the representation's redundancy. Their result matches human detection thresholds over many decades in light level. In constructing an approach they combine a design criterion, system constraints, and the measured statistics of images. These minimal ingredients allow for a parameter-free prediction, which implies that basic ideas of efficiency may have wide application in vision.

Laughlin [48] asks how a visual neuron should best encode contrasts so as to transmit as much information (in the Shannon measure [76]) as possible. The answer is to transform contrast in such a way that the response histogram is uniform. This procedure eliminates the same type of redundancy present in English text where letters of the alphabet are not used equally often. His treatment predicts a contrast-response curve which is quite similar to the response properties of LMC cells in a fly's visual system. Making this prediction requires the measured contrast histogram of natural images. We will return to this issue of response histograms later in the paper.

The basic idea in all of these approaches is that sensory systems are well-adapted for processing the types of signals present in nature. Many of the proposed critera for efficient design are based on statistical measures. Attneave [6] and Barlow [7] suggest that reducing the redundancy of sensory messages is a primary goal of sensory systems. Linsker has presented examples of ways in which sensory encodings can be maximally informative [55, 56, 57]. Other measures include reconstruction fidelity [57, 71], and the shape of neural response histograms [32, 48]. All of these criteria involve the statistics of the neural encoding. Only by first studying the properties of natural scenes can we make predictions as to how best to process them.

In summary, we often want to know how well an image processing system will work under normal conditions. One generally uses a statistical measure such as the average reconstruction error, of the amount of information the system delivers. This brings us into a framework which has a statistical basis and requires knowledge of natural image statistics.

3. Gathering natural images

Up to now we have been providing a motive for the study of natural images. Some important questions still remain:

- Which images should be in the ensemble?
- What should we measure?

- What statistics should we compute?

Clearly there is no 'right' ensemble. Each environment has its own typical characteristics and thus its own statistics. In interpreting the differences between these statistics we might seek creatures whose visual systems differ systematically with the statistics of the envronments they inhabit [38, 58].

One early study characterized television images [45] in an effort to understand how best to encode them. More recent work has focused on outdoor images of nature [31, 73, 82]. Our data consist of images of a wooded environment in springtime.

Once an environment is chosen we must decide which images to capture—that is, where to point the camera. We might also choose to use an angular resolution or spectral sensitivity which mimics a particular creature's visual system†. Of course visual systems function in real time and so analysing short movies instead of still images would add another level of detail.

Finally, we must consider the statistical analysis. Ultimately one would like to know $P[I]$, the probability of occurence of any image. But, as mentioned earlier, this would be impossible to achieve due to the required size of the dataset. Another possibility is to experimentally determine the best parameters for a model $P[I]$, such as a Markov random field [33, 43].

Perhaps the most direct approach is to catalog the correlation functions of the image distribution [13]. This means using image data to evaluate expressions like $\langle I(x_1)I(x_2)\cdots I(x_n)\rangle$. But with the exception of the second-order correlation function ($n = 2$), these quantities are difficult to interpret (and to visualize, since each spatial index adds two dimensions to the function's domain). The correlation functions enter naturally in a perturbative fashion into statistical calculations at low SNR, as discussed above. So although higher-order correlation functions may not provide much insight, they do have straightforward application.

Characterizing an arbitrary distribution can be done through brute force (by measuring correlation functions, for instance). But we can also seek a simple underlying structure or 'invariance' property in the distribution. That is, the image probability could transform simply if the image is transformed simply. One such symmetry is translation invariance. For some ensembles we expect that a given image appears with equal probability regardless of its positional offset. Such invariances can greatly reduce the complexity of the distribution. They are commonly sufficient to synthesize the distributions of quantum field theories [39], and also play an important role in image processing [51]. If the distribution of natural scenes contains no such invariance then just collecting statistics will be a useful—but not necessarily interesting—venture. However, we will find that natural images do indeed display some rather surprising symmetries.

Since there is no way to collect enough data to fully characterize an image environment, our statistical description will be far from complete. We will not even be able to reproduce realistic images with our minimal statistical knowledge. It is interesting to consider just how much knowledge of this kind is necessary to do something useful. For instance, we have seen that under conditions of low SNR the only important statistic is the power spectrum, and it can be used to design the optimal low SNR filter. At high SNR every detail of the statistics will matter, but how much? Could we limit our knowledge to a few statistics which allow a nearly optimal performance?

Effectively this is what our visual system does. The development of the mammalian visual system is strongly dependent on early visual stimulation [11, 61]. In particular, it

† Hopefully the characteristics of natural images will be fairly robust to these types of details.

is known that mammals raised in unusual image environments end up with functionally modified visual systems as adults [36]. This suggests that the visual system's development is influenced by the statistics of its environment, through some as yet unknown algorithm. Contained in the final 'wiring' of the visual system is a set of statistics about the creature's past visual experience. Knowing which statistics these are might provide great insight toward the nature of visual processing.

4. Image statistics

In 1952 Kretzmer [45] pioneered the modern analysis of real-world images. With applications to television image coding in mind, he tabulated a set of local image statistics such as the point histogram and the second-order correlation function. From these measurements he placed a lower bound on the image redundancy of about 3 bits per pixel.

The first mention of power-law scaling in image power spectra was in a 1957 paper by Deriugin [28], who also measured television signals. This property was rediscovered in 1978 by Cohen *et al* [23] (also see [20]), and again in 1987 by both Burton and Moorhead [16] and Field [31]. This scaling was later studied by Tolhurst *et al* [82] and by us [73, 74]. The ensemble power spectrum (averaged over orientations) is found to behave approximately as

$$S(k) \propto k^{-2+\eta} \tag{8}$$

where k is the modulus of the spatial frequency (in cycles deg^{-1}, for instance), and η is measured to be small. The power spectrum is a function of a single angular scale given by the spatial frequency, and it changes as a power of that scale. There is no preferred angular scale in natural images since the form of the power spectrum is invariant to any choice of basic scale. Doubling the spatial frequency always reduces the power by a factor of $2^{-2+\eta}$. This is not true, for instance, for a power spectrum of the form $S(k) \approx e^{-k/k_0}$, where k_0 acts as a 'typical' spatial frequency.

These studies provide evidence for a certain symmetry in ensembles of natural images: scale invariance. Scale invariance implies simply that the image statistics do not change with the angular scale. Pictures of such an ensemble will have the same ensemble statistics regardless of the lens' focal length. More generally, the new ensemble may be *self-affine* to the original one, meaning that the new images must also be multiplied by a suitable constant after rescaling to make the statistics identical to the original ones. If $Q[\phi(\alpha x)]$ is any ensemble statistic of $\phi(x)$ on scale α, then scale invariance implies that

$$Q[\phi(x)] = Q[\alpha^{\nu}\phi(\alpha x)] \tag{9}$$

where ν is a universal exponent (i.e. it is independent of both α and Q). Thus in a scale-invariant ensemble we can make the replacement $\phi(x) \to \alpha^{\nu}\phi(\alpha x)$ for all instances of ϕ in any expectation value.

This is a strong statement. It greatly restricts the form of the image distribution. Such a property also gives us some intuition about natural scenes instead of a mere quantification of their statistics. For instance, it reinforces the notion that objects in the natural world can appear at any angular scale in an image (i.e. they can be any distance away), which is one plausible mechanism for producing scale invariance.

Scale invariance is a widely studied property of critical phenomena, such as the Curie point of ferromagnets. Physicists study models with local (Markovian) interactions which give rise to 'long-range' (i.e. power-law) correlations. In order to attain scale-invariance, the model parameters must be chosen very precisely [89]. Interestingly, in two dimensions local scale- and rotationally-invariant models must also be conformally invariant [19]. But since

images in nature are not even isotropic—the horizon has a definite orientation—we cannot expect them to be more generally conformally invariant. Thus to model the scale-invariant statistics while excluding conformal invariance, one must include distant interations.

If natural images form a scale invariant ensemble, then we should find scaling in other statistics besides the power spectrum. In 1992 William Bialek and I collected natural image data and began a statistical analysis [70, 73, 74]. Aside from noted exceptions, the work presented below resulted from this collaboration. We were able not only to confirm the scaling result, but also to find evidence for a novel invariance.

Other work on natural images has included local principal components analysis [34, 52, 75], in which the local linear filters which maximally decorrelate the images are sought. The filters with highest response variance tend to have a resemblance to the oriented receptive fields found in the cortex, which suggests that some decorrelation process may be in operation. Decorrelation is commonly considered to be the first step in reducing the redundancy of a representation.

The spectral reflectance properties of natural scenes have also been studied. In 1947 Krinov [46] measured the spectral reflectances of 337 natural objects, such as grasses and wood, at 26 wavelengths. Maloney found that all these spectral profiles could be fit closely using models with 7 free parameters [59], which represents a redundancy in 19 of the 26 dimensions. Dannemiller [26] found that the noise due to random photon catches effectively reduces the dimensionality to 3, which is the number of cone types present in the human retina. Using colourimetric measurements (and thus invoking human colour vision specifically), Burton and Moorehead [16] found that natural images evoke highly correlated responses in cones, and showed that power spectra scale approximately as $1/k^2$ in each of the three cone systems. By studying properties of colour images one might try to predict the optimum arrangement of retinal photoreceptors [60], colour coding [3], and the optimal pupil function for chromatic vision [49].

5. Measuring natural images

As an image ensemble we chose a wooded environment in central New Jersey during springtime. These woods are the habitat of insects, small mammals, and birds, so it is an important sensory environment for many different types of visual systems. Since it constitutes only particular environment our results will not necessarily be characteristic of others. Another approach would be to gather images from widely varying environments and analyse them together as a single grand ensemble of natural scenes.

The woods consist of trees, scrub, rocks, and a stream. An image from the ensemble is presented in figure 1. Translation invariance is built into this ensemble since the camera was pointed in random directions with small elevation angle, and images of sky or ground alone were avoided. Details of the measurement process are presented in the appendix.

The luminance of a reflecting scene is proportional to the radiant flux from the sun. Visual systems, including cameras, adapt to this mean background value, making it irrelevant. It is removed from our images by considering logarithmic intensity fluctuations from a background level. We define the the 'log-contrast' $\phi(x)$ to be

$$\phi(x) = \ln\left[I(x)/I_0\right] \tag{10}$$

where $I(x)$ is the measured intensity signal, and I_0 is defined for each image such that $\sum_x \phi(x) = 0$. This gives every image histogram zero mean. The definition of I_0 is arbitrary, but many of our statistics are log-contrast differences in which the constant I_0 drops out. The logarithmic measure is convenient in that it covers the whole real axis;

Figure 1. Image from the woods: rocks in a stream with background foliage.

intensities, on the other hand, are somewhat difficult to work with since they are non-negative. We find the use of ϕ instead of I also seems to improve the observed invariances.

The data set consists of 45 images taken at a 15 mm focal length (1 pixel subtends 0.059° of visual angle) and 25 images at an 80 mm focal length (0.011° per pixel). The digitized images are 640×480 pixels, from which the central 256×256 subimage is taken. At 15 mm and 80 mm focal lengths the images subtend 15° and 2.8° respectively.

6. Statistical analysis

Previous work has hinted at scaling in natural images. We will try first to substantiate it by evaluating the power spectrum of our images. Digitized images are discretely sampled continuous images, $\phi(x)$, which are drawn from a stationary distribution with power spectrum $\mathcal{S}(k)$ given by

$$\langle \Phi^*(k)\Phi(k') \rangle = (2\pi)^2 \mathcal{S}(k)\delta^{(2)}(k - k') \tag{11}$$

where $\Phi(k) = \int d^2x\ \phi(x)e^{-ik\cdot x}$. The coordinate system, strictly speaking, consists of the azimuth and elevation angles. The pixelated image is derived from the continuous image as

$$\phi_{m,n} = \phi(ma, na) \tag{12}$$

where a is the pixel spacing measured in degrees, and m, n are pixel indices which go from 0 to 255. We want to estimate $\mathcal{S}(k)$ from the discretely sampled images.

Spectral estimation is a well explored field, a cogent summary of which can be found in [65]. The power spectrum is estimated using

$$S_{\text{est}}(k_{r,s}) = \left(\frac{a}{2\pi M^2}\right)^2 \frac{1}{N}\sum_{i=1}^{N}\left|\sum_{m,n=0}^{M-1}\phi^i_{m,n}W_{m,n}\exp\left(-ik_{r,s}\cdot x_{m,n}\right)\right|^2 \tag{13}$$

where $x_{m,n} = (ma, na)$, $k_{r,s} = (2\pi r/Ma, 2\pi s/Ma)$, $M = 256$, and i runs over all the images in the data set. $W_{m,n}$ is called the 'windowing function'; its shape determines how the estimate relates to the actual spectrum. We use a two-dimensional Bartlett window in our computations, but the result is not strongly dependent on this choice.

A contour plot of the power spectrum of the natural scene log-contrast is presented in figure 2. The contours are placed at constant intervals in the logarithm of the power. It shows a preponderance of power in low spatial frequencies along the horizontal and vertical orientations, which are clearly special. It is more illuminating to plot the orientationally averaged spectrum, which is shown in figure 3. The plot consists of two superposed graphs, each from a different lens focal length. The scale is logarithmic both in spatial frequency and power, and the data thus plotted are nearly linear. This means that the power spectrum is a power-law of the form

$$\mathcal{S}(|k|) = \frac{A}{|k|^{2-\eta}} \tag{14}$$

with $\eta = 0.19 \pm 0.01$, and $A = (6.47 \pm 0.13) \times 10^{-3}$ deg$^{(0.19)}$.

For a given focal length measurement difficulties arise at high spatial frequencies. Optical blur causes the spectrum to fall, and at the same time noise and aliasing cause an increase in power. We extend the spatial frequency range by simply using two focal lengths. The graph shows scaling of the spectrum over nearly 2.5 orders of magnitude in spatial frequency. The highest frequency for which the scaling is demonstrated is about 30 cycles deg^{-1}, which corresponds to about half the acuity of the human eye [17]. These are thus relevant spatial frequencies for vision. Although our results derive from the logarithm of intensity, we find the same scaling in the power spectrum that others have found when using the intensity signal (we find this as well). Scaling in the power spectrum is robust to such a change, as well as to the obvious differences in choice of ensemble, spectral sensitivity, and methods of image capture. However, the exponent differs between authors and environments (e.g. we find $\eta \approx -0.3$ for beach images).

This scaling of the power spectrum confirms the findings of others. But we can say more. First, the power spectrum alone does not tell us whether the distribution is Gaussian. Also, scaling should be testable through any statistic of our choosing. Both issues can be explored at the same time through the process of 'coarse-graining.' The original scaling ideas of Kadanoff [44] propose that a coarse-grained critical system should have the same statistics as the original system, aside from a possible rescaling of the field variables.

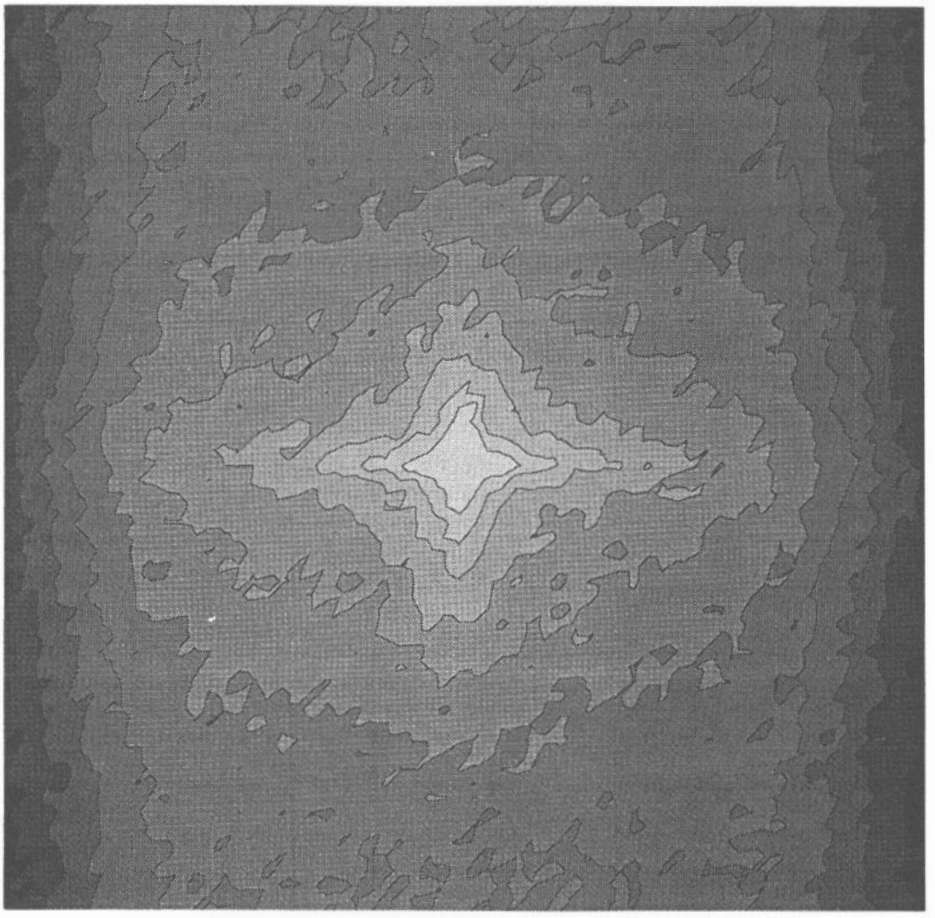

Figure 2. Contour plot of ensemble power spectrum of 45 images taken at focal length of 15 mm. Center of figure is $k = 0$. Contours are placed at equal intervals in the logarithm of power, and spatial frequency is plotted on linear axes.

Similarly, we can coarse-grain the images to look for scaling and the possibility of Gaussian statistics. The average of ϕ over scale N is given by

$$\phi_N = \frac{1}{N^2} \sum_{m,n=1}^{N} \phi_{m,n}. \tag{15}$$

An example of this procedure is shown for $N = 2$ in figure 4.

If ϕ is a scaling field then the probability $P_N(\phi_N)$ should have a *shape* which is independent of N. In the theory of critical phenomena, when a field has an anomalous dimension (i.e. $\eta \neq 0$) it must be 'renormalized' when length scales are changed. This implies

$$P_N(\phi) = 1/\phi_N^{\mathrm{RMS}}\, \mathcal{P}(\phi/\phi_N^{\mathrm{RMS}}) \tag{16}$$

where $\phi_N^{\mathrm{RMS}} = \langle \phi_N^2 \rangle^{1/2}$, and $\mathcal{P}$ is the scaling probability distribution. For a scaling field with anomalous dimension $\eta/2$, $\phi_N^{\mathrm{RMS}} \propto N^{-\eta/2}$. In order to compare histogram shapes this RMS value must be divided out.

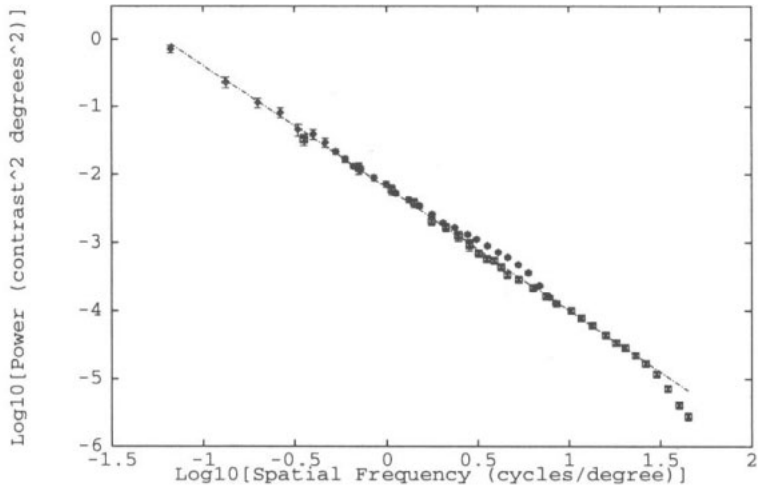

Figure 3. Orientationally averaged power spectrum with standard error bars for 15 mm and 80 mm focal length data (overlapping), along with the regression line fitted as discussed in the text.

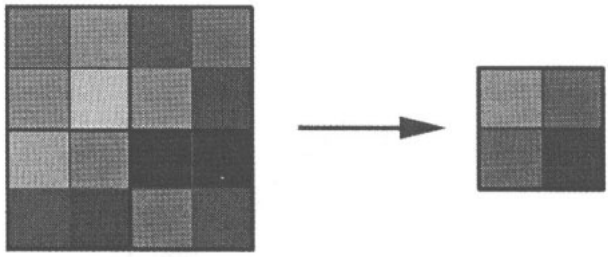

Figure 4. Block averaging procedure for $N = 2$.

If $\phi(x)$ scales then the probability distribution of ϕ_N should always have the same shape, regardless of the value of the rescaling parameter N. Figure 5 demonstrates scaling of the log-contrast distribution. The plot shows $P(\phi_N/\phi_N^{\mathrm{RMS}})$ versus $\phi_N/\phi_N^{\mathrm{RMS}}$ for $N = 1$, 2, 4, 8, 16, and 32. These six graphs all lie on top of one another; they have the same shape. A Gaussian distribution would show a parabola in the plots instead of the nearly linear tails we find. Histogram scaling is a much stronger statement than the scaling of a two-point function (i.e. the power spectrum), since it means that higher-order correlation functions must also scale.

The central limit theorem states that when a large number of independent random variables (with finite variance) are averaged together the resulting distribution becomes Gaussian. The fact that our histograms do not become Gaussian even after averaging together nearly 1000 (32×32) data points is evidence for what physicists call a non-Gaussian scaling fixed point. Simply put, the central limit theorem does not apply since the variables being averaged are highly correlated. Such extended correlations are typical of a thermodynamic critical point, where the correlation length is infinite.

As another example of both the scaling of statistics and the non-Gaussian character of the ensemble, consider the distribution of local gradients. We first block the images on a

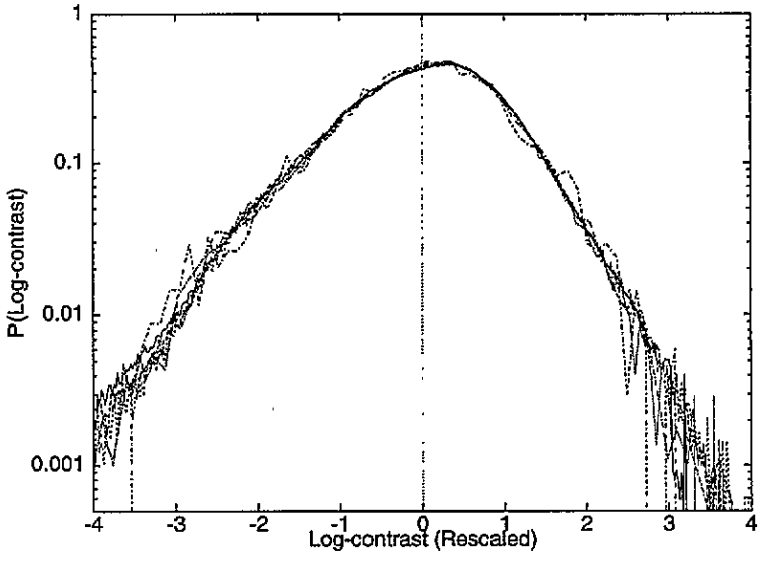

Figure 5. Scaling of log-contrast histograms over scales 1, 2, 4, 8, 16, and 32.

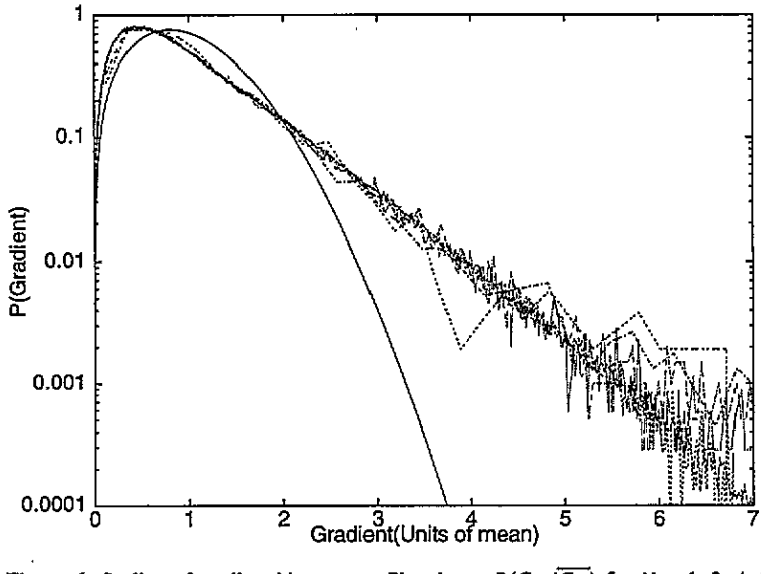

Figure 6. Scaling of gradient histograms. Plot shows $P(G_N/\overline{G_N})$ for $N = 1, 2, 4, 8, 16, 32$ with Rayleigh distribution (solid) shown for comparison.

scale N and then calculate a discrete approximation to the magnitude of the gradient,

$$G_N(m, n) \approx |\nabla \phi_N(m, n)|. \tag{17}$$

If ϕ is a scaling field then the histogram of G should have a shape which is independent of N. Figure 6 shows $P(G_N/\overline{G_N})$ versus $G_N/\overline{G_N}$, where $\overline{G_N}$ is the mean of the distribution. If ϕ were Gaussian, then this distribution would take the Rayleigh form

$$P(G) = \frac{\pi}{2}\frac{G}{\overline{G}^2}\exp\left[-\frac{\pi}{4}\left(\frac{G}{\overline{G}}\right)^2\right] \tag{18}$$

which is plotted in the figure for comparison.

The distributions of gradients for length scales from 1 to 32 are identical in shape over nearly four decades in probability. There is a stark contrast between this distribution and the Rayleigh form. First, the histogram of gradients has a very long exponential tail where the Rayleigh distribution falls off much more sharply. This means there are far more regions of large gradient in the images than there would be if they were Gaussian. Also, there is an excess of small gradients, or uniform patches. These are non-Gaussian signatures of natural scenes. Such patterns of very large and very small gradients are seen in thermally driven convective turbulence, which also gives rise to non-Gaussian probability distributions with exponential tails [21], and displays scaling [67].

Scaling allows us to replace $\phi(x)$ by $\alpha^{\nu}\phi(\alpha x)$ without changing any of the statistics. We have seen that rescaling images does not change the shape of the log-contrast distribution, only its width ϕ_N^{RMS}. According to the scaling law this width should scale as $N^{-\nu}$. In figure 7 we plot ϕ_N^{RMS} versus N on a log–log scale. The graph is linear with slope $-\nu \approx -0.2$. All the local quantities we have tested scale with this same exponent†. Had the pixels been independent of one another, the rescaling factor would have fallen as N^{-1}; this line is plotted for comparison. The variance in natural images remains characteristically larger than it would for white noise when images are averaged over large scales.

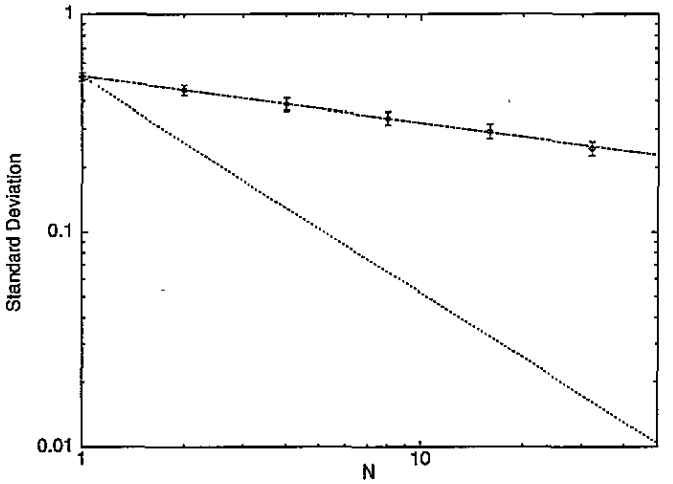

Figure 7. Standard deviation of ϕ_N distribution as a function of N (log–log plot). If the images were made of uncorrelated noise, the standard deviation would scale as $1/N$, as shown by the dotted line.

7. Predictability in natural images

One measure of the non-randomness of images is the amount of predictability they contain. Suppose we know what some sections of an image looked like, and from them we want to guess what the missing sections are. How well can this be done? The answer lies, as

† The astute reader may wonder why $\eta \neq 2\nu$, since the power spectrum involves two powers of the field ϕ. This does present something of a mystery, but it can possibly be resolved by considering the fact that the scaling may not be perfect and isotropic. The way orientations were averaged when computing the power spectrum is different from the way orientations are confounded when averaging over square blocks, which might explain the difference in exponents.

one might expect, in the statistical structure of the images. If they are composed of random pixels, then there is no predictability. But natural images possess long-range correlations, and so a large degree of predictability is expected.

Claude Shannon, the inventor of information theory, held an interest in statistics of the English language. In a 1951 paper [77] he used the inherent knowledge of native speakers to place bounds on the redundancy (or predictability) of written English. He would remove the last character of a string of n characters from an English text, and the native speaker would fill in the missing character in as few tries as possible. From the histogram of the number of guesses until a correct response Shannon placed bounds of between 0.6 and 1.3 bits of entropy per letter for $n = 100$. This represents a single letter redundancy of about 75%.

In 1987 Kersten asked human subjects to perform a similar task on everyday images. A pixel was removed from an image and a subject was asked to replace it. Using the method of Shannon, he placed the single pixel redundancy in natural scenes at 65%. This means that of all the entropy a pixel has, 65% of it is predictable from knowledge of the rest of the image.

In a collaboration with Horace Barlow and Chris Wroe in Cambridge, we used the natural images to make some assessments of predictability. One would ultimately like to know how much information a patch of image conveys about another patch some distance away. Such a computation would involve an immense ensemble of images so as to sample the distribution well. The most that can be practically accomplished is to ask about a few pixels at a time. For example, how much information does one pixel convey about another a given distance away?

We compute Shannon's mutual information between the two pixel values, ϕ_1 and ϕ_2

$$I = \int \mathrm{d}\phi_1 \, \mathrm{d}\phi_2 \; p_2(\phi_1, \phi_2) \log_2 \frac{p_2(\phi_1, \phi_2)}{p_1(\phi_1) p_1(\phi_2)} \tag{19}$$

where p_2 is the joint distribution of two pixels at a given separation, and p_1 is the marginal distribution of a pixel. For a given displacement vector d all positions in the images are scanned to create a histogram of joint probabilities. From this distribution a discrete approximation to equation (19) is computed.

The left graph of figure 8 shows the mutual information between two pixels as a function of the distance, d, between vertically separated pixels. The graph is very nearly linear on a log–log scale, meaning the information scales as a power-law in the separation distance:

$$I(d, \theta) \approx d^{-\alpha(\theta)} \tag{20}$$

where $\alpha(\theta)$ conveys the dependence of the slope on the angle of the separation axis. The right graph shows the systematic anisotropy in α as a function of θ ($\theta = 0$ corresponds to a horizontal separation). The fact that the slope is smallest for $\theta = 90°$ implies that correlations are strongest vertically, as one might expect in images containing trees. We compute the single pixel nearest-neighbour redundancy to be about 10%. Kersten's much larger figure reflects the fact that the redundancy is present in substantially larger areas of the image.

Interestingly, a similar scaling of information is found in English texts. Ebeling and Pöschel have found that the mutual information between two letters as a function of distance scales as $I \approx d^{-0.37}$ up to distances of about 100 letters, where finite data-set noise began to creep into the measurement [30].

Note that the scaling of information with distance is quite fortuitous, as such a complicated statistic has no *a priori* reason to be scale-invariant, even if the image ensemble is. In a scale-invariant ensemble any single-power correlation function will scale,

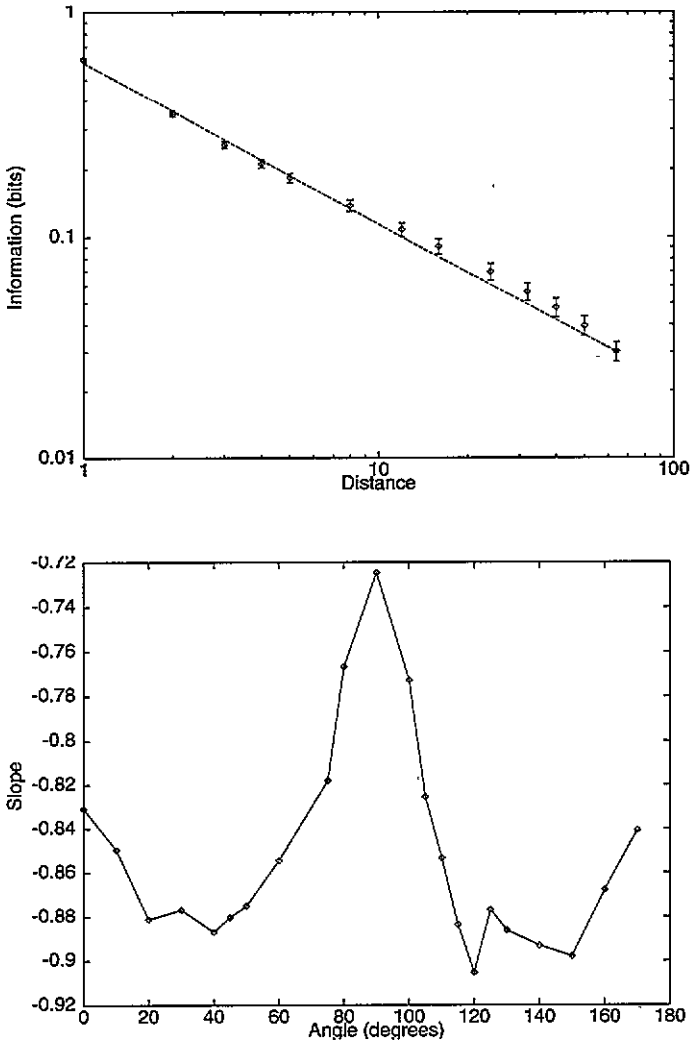

Figure 8. Top: log–log plot of $I(d,\theta)$ versus d for $\theta = 90°$. Bottom: scaling exponents of $I(d,\theta)$ versus θ.

such as $\langle\phi^2(0)\phi^4(x)\rangle$. But this is not necessarily true of a mixture of powers, such as $\langle(\phi(0)\phi(x)+\phi^2(0)\phi^2(x))\rangle$, since a change in length scale rescales each term by a different exponent (see equation (9)), and so the function changes. The information measure above is a special quantity which does happen to scale even though it is not a single power of the scaling field ϕ.

We can try another measure of correlation, namely the second-order correlation coefficient, given by

$$r = \frac{\langle\phi_1\phi_2\rangle}{\sqrt{\langle\phi_1^2\rangle\langle\phi_2^2\rangle}} \tag{21}$$

assuming $\langle\phi_1\rangle = \langle\phi_2\rangle = 0$. This statistic turns out not to obey a power law; instead it crosses over from power-law to exponential behaviour at distances larger than about 20 pixels. Since

scaling is a basic statistic of natural images, Shannon's mutual information may be a more 'fundamental' measure of correlation than the second-order correlation coefficient.

As a final example of two-point image statistics we examine the degree of reconstructability that one pixel provides about another. How well can ϕ_2 be estimated from a pixel ϕ_1 lying certain distance away? We want a function $\phi_2^{\text{est}}(\phi_1)$ which is the best guess for ϕ_2 given ϕ_1. The estimate which minimizes the mean-squared error, $\langle|\phi_2 - \phi_2^{\text{est}}(\phi_1)|^2\rangle$, is

$$\phi_2^{\text{est}}(\phi_1) = \int \mathrm{d}\phi_2 \, \phi_2 p(\phi_2|\phi_1) \tag{22}$$

which is the conditional mean of ϕ_2 given ϕ_1.

Figure 9 shows $\phi^{\text{est}}(\phi)$ for vertically separated pixels at distances of 1, 4, and 16 pixels. The relative importance of different regions of the plot is indicated by an overlayed graph of the pixel probabilities. In the relevant region the estimate is a linear prediction, i.e.

$$\phi_2^{\text{est}}(\phi_1) \approx m(d)\phi_1 \tag{23}$$

where $m(d)$ is the slope of the line as a function of the separation distance between the two pixels. Figure 10 shows the functions $m(d)$ and $F(d)$, the reconstruction fidelity defined as

$$F = 1 - \frac{\langle|\phi^{\text{est}} - \phi|^2\rangle}{\langle\phi^2\rangle} \tag{24}$$

Both are power-law in form, with exponents -0.31 and -0.72, respectively.

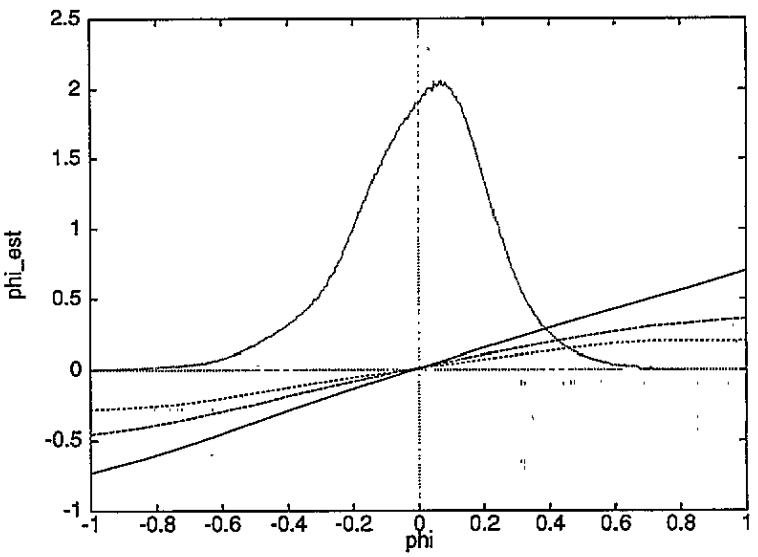

Figure 9. Estimator $\phi^{\text{est}}(\phi)$ for pixels separated vertically by 1 (solid), 4 (dashed), and 16 (dotted) pixels. The probability distribution of ϕ is plotted to show the region of interest.

Reconstruction based on the knowledge of a single pixel is not very good. The RMS prediction error based on knowing a vertical neighbour is nearly 70% of the RMS pixel variations themselves (i.e. the error we would get if we had no such knowledge). The estimates can improve by using more of the nearby pixels in the form of an optimal estimator. As more pixels are included the optimal estimator takes into account more and more of the image statistics in the form of joint pixel densities. Nothing particularly special about natural images is captured in the two-point pixel statistics.

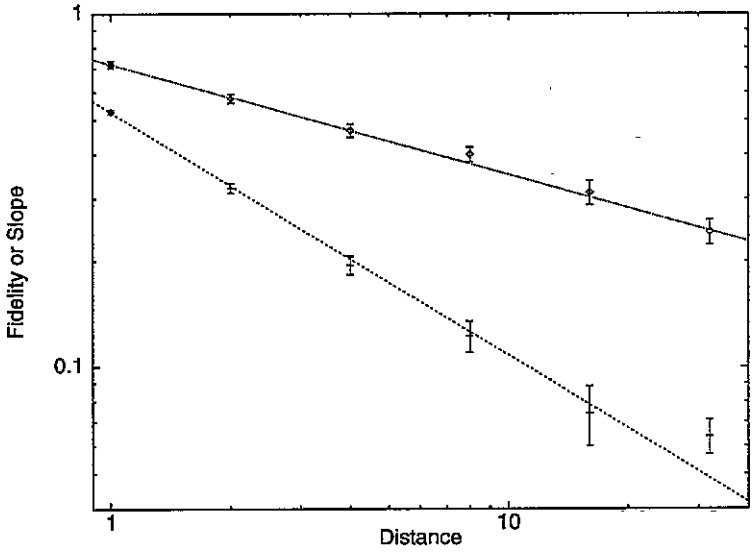

Figure 10. Reconstruction fidelity F (dashed), and slope of linear reconstruction m (dotted) as a function of distance d (log–log).

Even without the complete pixel statistics we can find the optimal linear estimator of a pixel from its nearest neighbours (NN)

$$\phi^{\text{est}} = \sum_{i \in \text{NN}} a_i \, \phi_i \, . \tag{25}$$

It requires only second-order correlations. The coefficient vector a is calculated to minimize the expected mean-square error as

$$a = C^{-1} y \tag{26}$$

where $C_{ij} = \langle \phi_i \phi_j \rangle$, $y_i = \langle \phi \phi_i \rangle$, and the indices i and j run over the 8 nearest neighbours. The coefficients a_i are slightly negative along the diagonals and have values of about 0.3 along the horizontal and vertical. This reconstruction provides an RMS error of about 50% of the RMS pixel fluctuations. A reconstruction which includes next-nearest neighbours brings the error down to 45%.

8. A new invariance

8.1. Filtering natural images

The images we have been exploring are examples of the signals which the visual system processes. What do their statistics tell us about how this processing should be done? The early stages of vision, such as those in the retina, are constrained to process images locally—no neuron has access to the entire image [29]. The neurons which convey these signals will have output statistics which are determined by the images. According to various efficiency criteria the responses of these channels should have certain statistical properties.

For instance, channels with signal variance constraints are optimized for information transfer by sending Gaussian signals [76]. Neurons have an analogous constraint in their function since their firing rates can saturate at high levels and cannot go negative. The optimal encoding statistics thus depend on the imposed constraints.

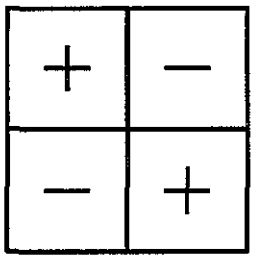

Figure 11. A local bandpass filter.

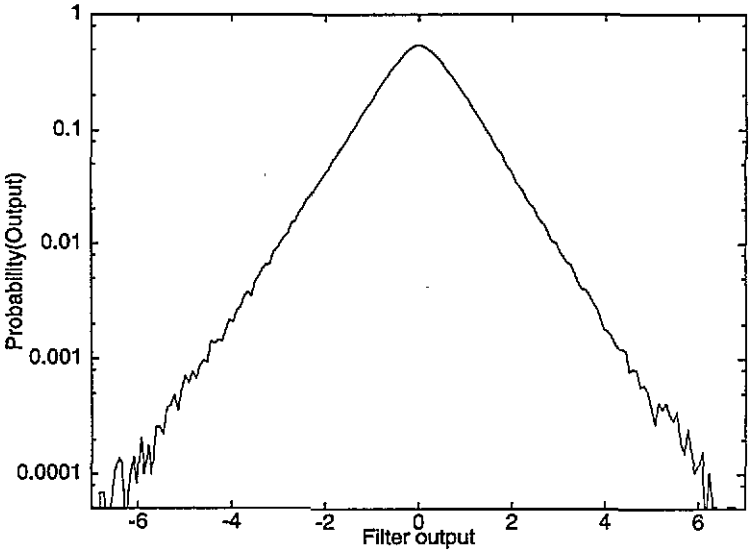

Figure 12. Histogram of the output of a local filter.

First we should determine the types of statistics which come from simple linear filtering of the log-contrasts (filtering the intensities gives similar results). Consider a filter of the form shown in figure 11. This 2×2 filter passes no signal at zero spatial frequency, and thus I_0 drops out. The histogram of this filter's output is shown in figure 12 on a semi-log scale. It has nearly perfect exponential tails over four decades in probability. In fact *any* local linear transformation we try (including Gabor and centre-surround filters) seems to produce exponential-tailed histograms, though their shapes can differ somewhat.

Long-tailed histograms from natural scenes have also been seen by Daugman [27], Barlow and Tolhurst [9], and Field [32]. Field points out that such 'sparse coding' is a consequence of specific arrangements in the Fourier phases of natural images. He proposes that long tails have the effect of activating only a very sparse subset of the neurons which code images. Burr and Morrone [15] believe that this property of images is related to the existence of edges in natural scenes, and 'signals features of interest to vision.' Images drawn from a Gaussian distribution have completely random phases, and show no structural resemblance to natural scenes [32, 70].

A one-sided exponential distribution maximizes information transmission for a given mean activation level, just as a Gaussian is optimal for fixed variance. If the neurons encoding visual stimuli have a mean firing rate constraint, then these exponential histograms are ideal; we just need to rectify them so that they are one-sided.

Is there a way to transform away the exponential distributions? No local linear transformation does it. But what about local *nonlinear* transformations? One method is simply to find a pointwise transformation of the image which produces the distribution we want, as Laughlin does to predict contrast coding [48]. But his method does not find the cause of the histogram, it just re-shapes it. We prefer to seek the mechanism which produces the long tails and systematically 'undoes' it.

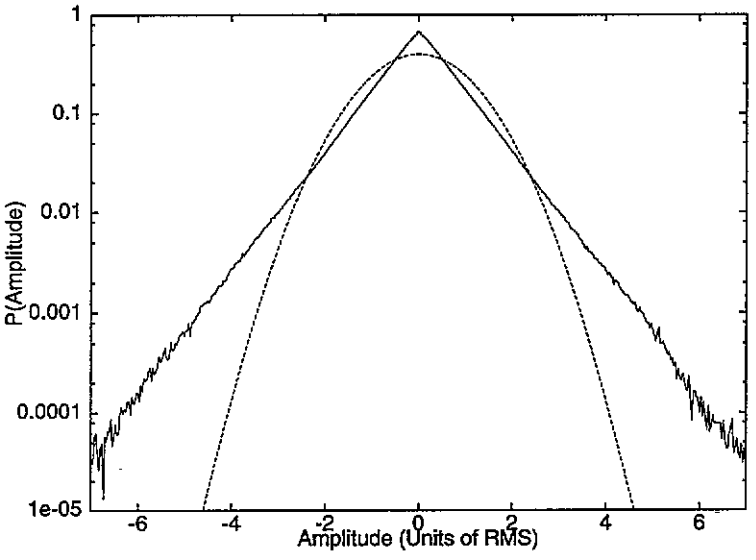

Figure 13. Histogram of amplitudes of a 5 min sequence from Strauss' 'The Blue Danube' from a compact disk recording as sampled at 44 kHz using a linear analogue-to-digital convertor (ADC) with 16-bit resolution. A Gaussian distribution is shown to highlight the excesses in the peak and the tails of the histogram.

8.2. Adding a nonlinearity

To search for a likely candidate we should think about the possible causes of the excess histogram tails. Consider an analogy with music, which is an ensemble with similar properties to images. The amplitudes of musical sound pressure also have exponential tails†‡ (see figure 13). The source of the long tails is the *dynamics* of the musical score; some sections are loud and some are quiet for an interval of time. If the quiet passages were amplified and the loud ones attenuated then the excesses at the tails and the peak of the distribution would move to more 'typical' values, thus diminishing the peak and tails. This would give the distribution a more 'rounded' or Gaussian character. Maybe a similar dynamic occurs in natural scenes, where locally correlated regions are either flat in texture (i.e. quiet) or very dynamic (loud). This suggests an origin for long exponential tails: The histogram is a superposition of many distributions of different variance.

Dividing a sound waveform by its recent loudness is a local nonlinear operation. We can try an analogous procedure on images by normalizing log-contrast fluctuations relative

† Exponential tails in histograms of real-world phenomena may in fact be quite general [87].

‡ It is also well known that many forms of Western music have scale-invariant statistics [37, 85].

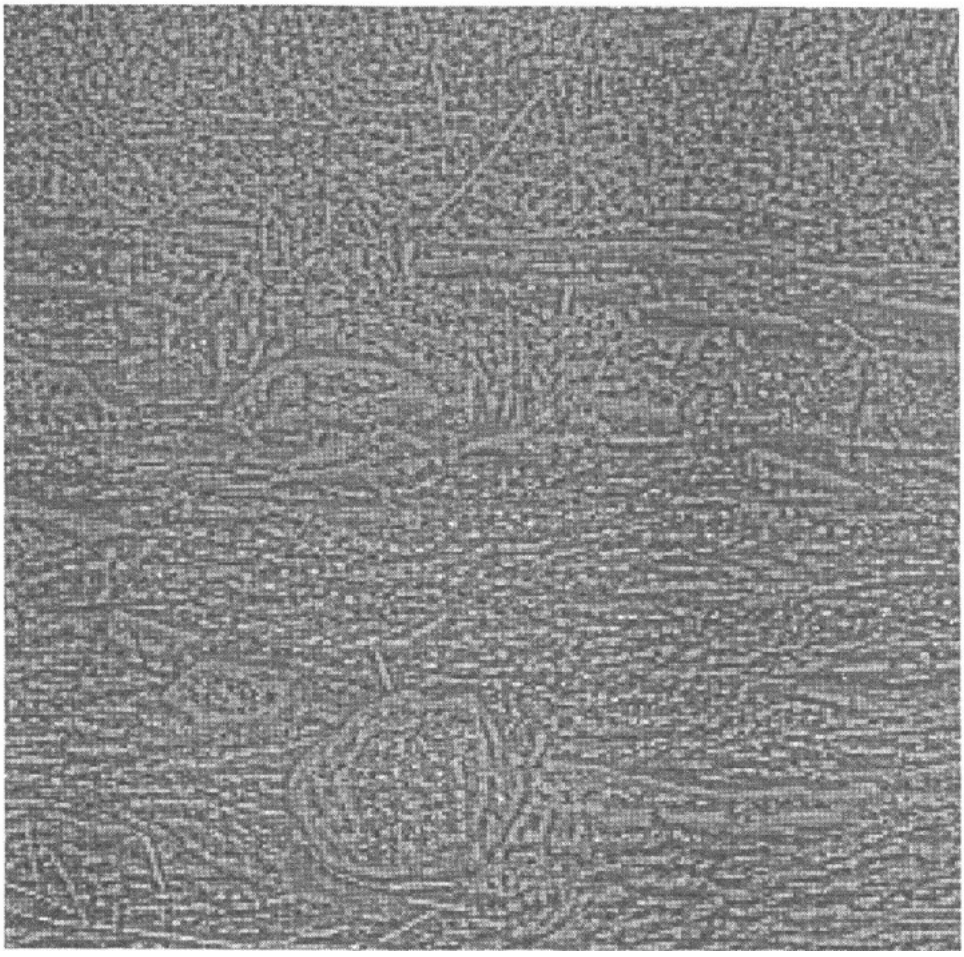

Figure 14. Variance modified image, $\psi(x)$.

to their local standard deviation. This creates a new field

$$\psi(\boldsymbol{x}) = \frac{\phi(\boldsymbol{x}) - \overline{\phi}(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \tag{27}$$

where $\overline{\phi}(\boldsymbol{x})$ is the local mean within the $N \times N$ block surrounding position $\boldsymbol{x}$, and $\sigma(\boldsymbol{x})$ is the standard deviation of ϕ within the block. This procedure has the effect of removing mean displacements from zero log-contrast and normalizing the local variance of the log-contrast. Patches of small local contrast will be expanded, and high contrast areas will be toned down. The numerator is like a centre-surround mechanism with the surround N times as large as the centre (1 pixel). Running the procedure on the image in figure 1 (using $N = 5$) gives a variance modified image, ψ, shown in figure 14, and a standard deviation image, σ, shown in figure 15. The variance normalized image is much more homogeneous than the original. Small fluctuations on the rock, for instance, have been expanded out to higher contrast. The image almost looks like a noise pattern, except for a few residual object borders. On the other hand, the standard deviation image, σ, seems

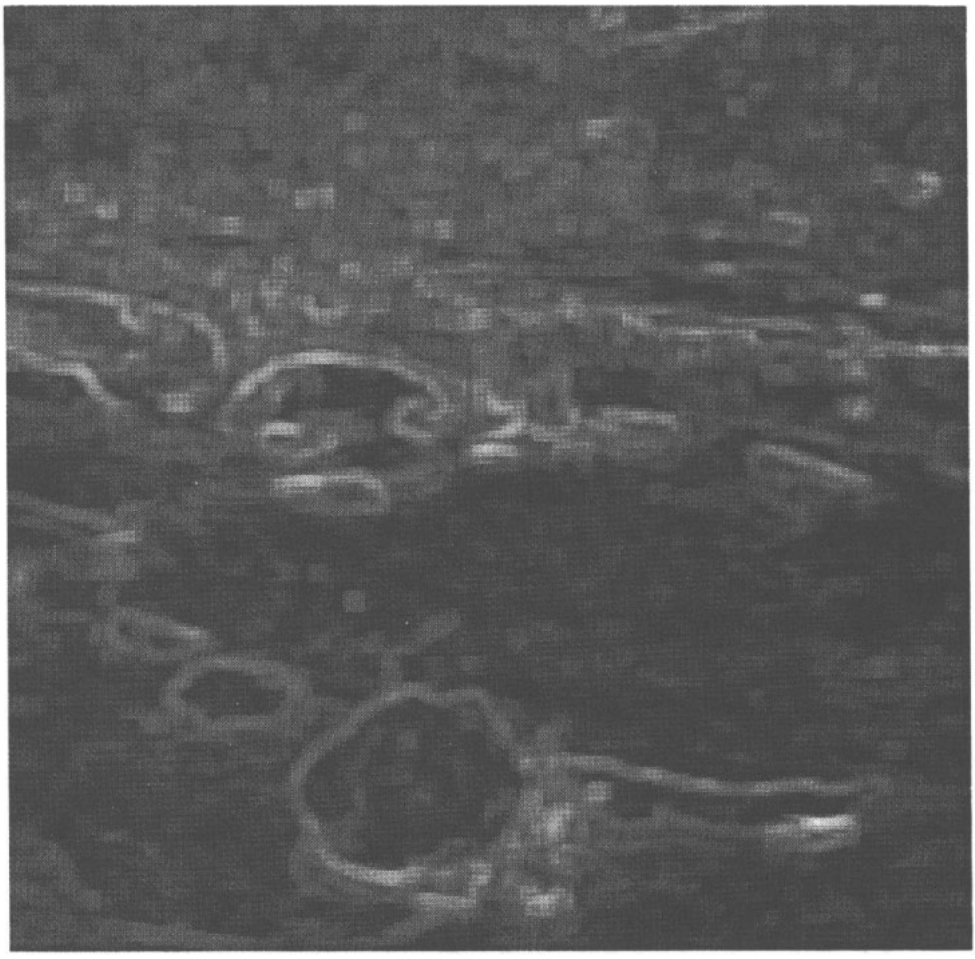

Figure 15. Standard deviation image, $\sigma(x)$.

almost to highlight the object borders and to attenuate the object textures. It is very roughly as if variance normalization separates objects from their textures.

We find that for the value $N = 5$, the resulting histogram of $\psi(x)$ is closest to a Gaussian, in that its kurtosis is nearest to that of a Gaussian ($3\sigma^4$). For smaller N the kurtosis is greater, and for larger N it is less. The statistics of ψ are shown in figure 16. The variance modified images are not exactly Gaussian, but they show Gaussian signatures: The histogram tails of ψ fall off rapidly, and its gradients are Rayleigh distributed. This new signal is thus amenable to transfer down a dynamically limited channel.

What about $\sigma(x)$? Making images from the variance, $\sigma^2(x)$, averaged and sampled in 5×5 blocks, gives us a set of reduced size non-negative images. We can treat these in the same way as we did the original image data by looking at the log-contrast. Define

$$\xi(x) = \ln\left[\frac{\sigma^2(x)}{\sigma_0^2}\right] \tag{28}$$

where σ_0^2 is analogous to I_0. The statistics of ξ are very similar to those of ϕ, as shown

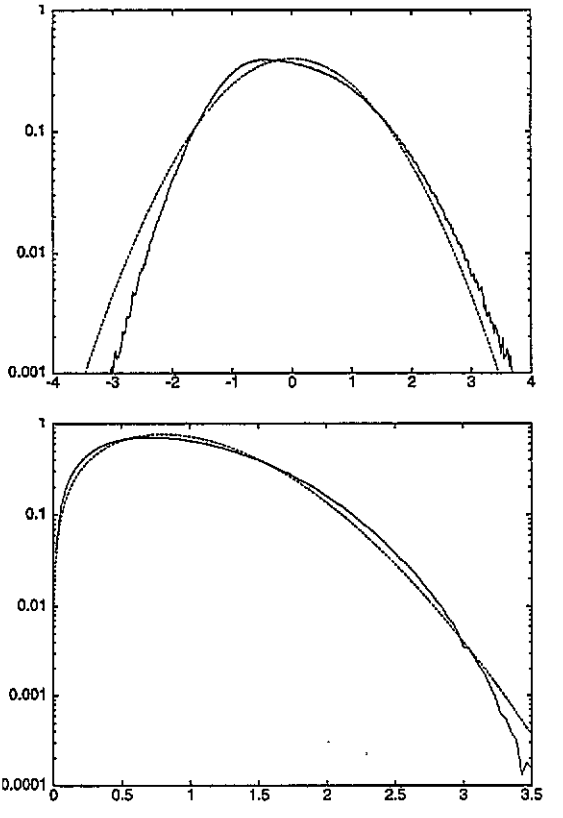

Figure 16. Semi-logarithmic plots of the statistics of variance modified images, ψ. Top: histogram of ψ (rescaled to unit variance) shown with a Gaussian for comparison . Bottom: histogram of gradients of ψ shown with a Rayleigh distribution for comparison .

in figure 17. Both the log-contrast and log-contrast gradient distributions of the original images and the variance images are quite similar. The power spectra are compared between ϕ and the variance images before subsampling, so they are on the same spatial frequency scale. The slopes of the spectra agree completely at low frequency (they have been shifted to match vertically). At high frequency the variance image spectrum falls off, as expected from the low-pass nature of the statistic.

8.3. Re-iterating the procedure

For every statistic we measure, the variance images are identical to the original ones. This means the patterns of local variances in natural images are statistically much like the patterns of intensity. Does this mean the entire procedure can be re-iterated? The answer is yes, but the results are not quite as clean as before. Start with the full resolution variance images and produce two datsets, $\zeta(x)$ and $\Sigma(x)$, which are the variance normalized and standard deviation images, respectively, of ξ (see figure 18). The most Gaussian ζ statistics are found by averaging and sampling in 11×11 blocks (see figure 19). This is the length scale at which the kurtosis of ζ crosses zero. Similarly, the distribution of gradients is closest to Rayleigh when ζ is sampled and averaged over 11×11 blocks; it is shown in figure 19. This procedure, iterated a second time, has produced another set of nearly Gaussian data plus a low resolution set of variances. How do these new variances compare with the original

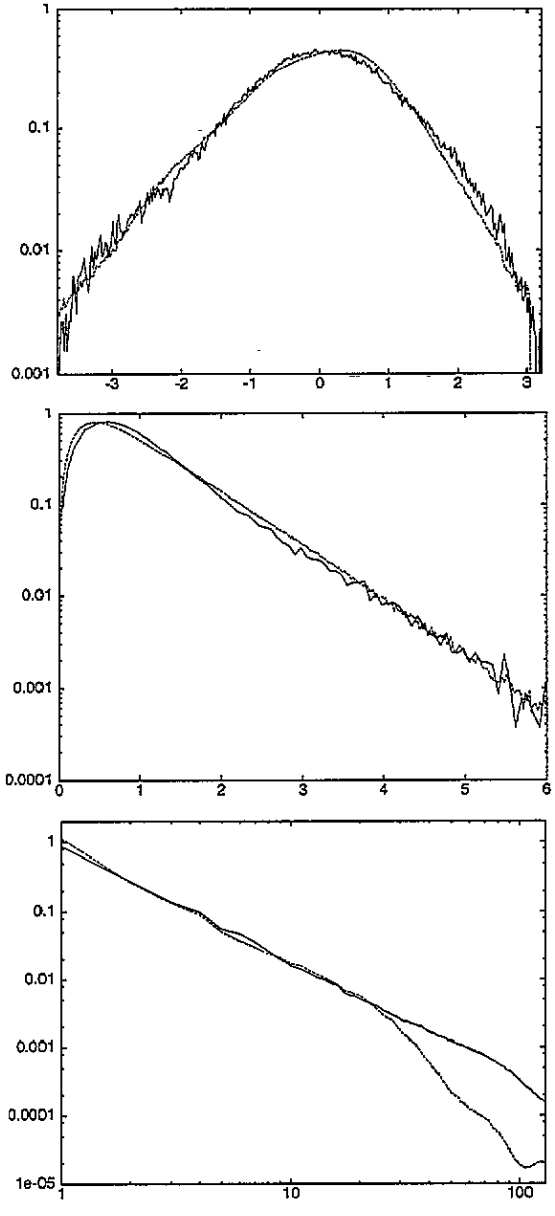

Figure 17. Statistics of ξ images compared to the original images, ϕ . Top: histogram of log-contrasts (scaled to unit variance, semi-log plot). Middle: histogram of gradients (scaled to unit mean, semi-log plot). Bottom: power spectra of ξ images (falls off at high frequency) and original images (arbitrary spatial frequency units, log–log plot).

images? Define a new log-contrast:

$$z(x) = \ln\left[\frac{\Sigma^2(x)}{\Sigma_0^2}\right] . \tag{29}$$

Its statistics (averaged over 11×11 blocks) are shown in figure 20, along with those of the original image log-contrasts. Again, the match is nearly perfect. Note that this was not a

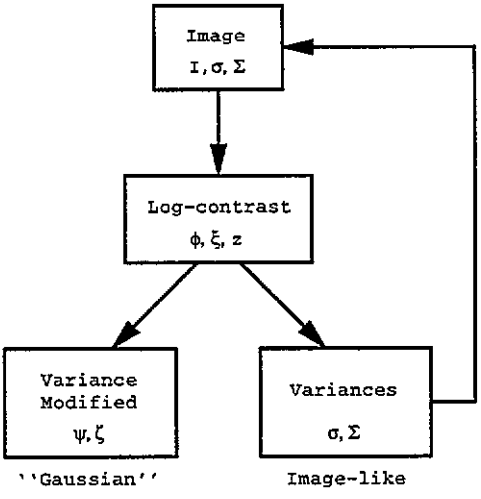

Figure 18. The iterated variance normalization procedure.

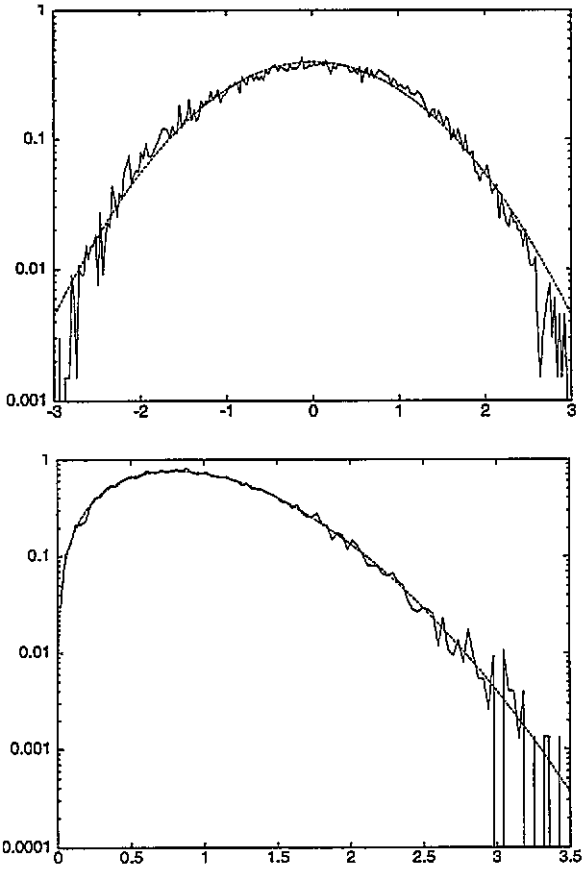

Figure 19. Statistics of ζ images compared with those of a Gaussian field . Top: histogram of ζ (rescaled to unit variance, semi-log plot). Bottom: histogram of gradients of ζ (rescaled to unit mean, semi-log plot).

true re-iteration of the procedure, as that would have meant averaging over 5×5 blocks in reduced images of σ. But this latter method does not reproduce the statistics quite as well.

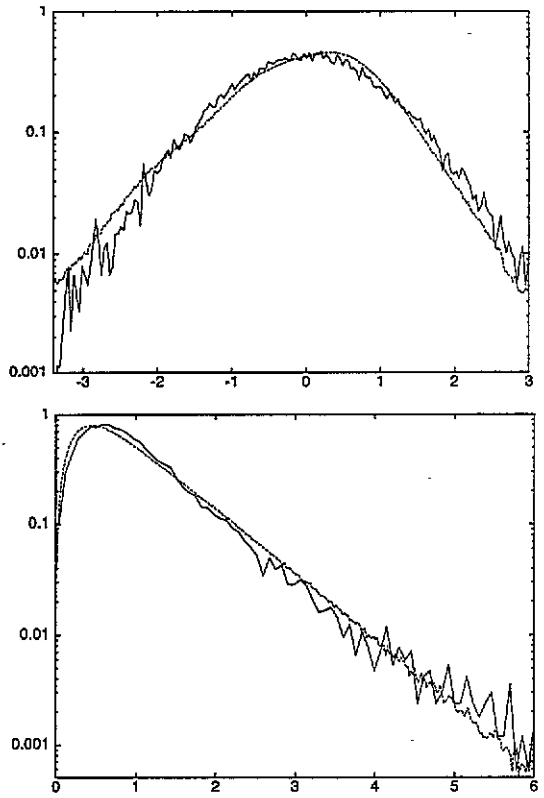

Figure 20. Statistics of z images compared with the original images, ϕ . Top: histogram of log-contrasts (rescaled to unit variance, semi-log plot). Bottom: histogram of gradients of z (rescaled to unit mean, semi-log plot).

Due to a lack of data, the procedure cannot be reasonably performed a third time. If it could be continued indefinitely then we have found a nonlinear invariance of the image ensemble. Local variances themselves have local variances with the same statistics. Thus correlations exist not only between image intensities, but also between local variances. The variance normalization procedure is analogous to spectral whitening in that it removes the correlation, though it is done hierarchically instead. The invariance has implications for coding: one could iterate the procedure to produce at each stage a set of Gaussian signals which could be communicated efficiently. In the end images could be made completely Gaussian. Most importantly, this has all been accomplished with local processing. These results may also imply that naturalistic scenes could be generated from Gaussian noise by inverting this procedure. Accomplishing this would mean that the structure of $P[\phi]$ has been completely understood.

In short, we have discovered a new type of iterative invariance in natural scenes. This variance normalization procedure is remeniscent of the 'contrast gain control' mechanism found in the visual cortex [35, 63]. It may simply serve to limit the variance of neural responses by reducing the tails of their distribution. Iterating the procedure reproduces the same statistics again and again, which implies that a universal algorithm may serve ideally

to process these image components.

Note that the procedure described above is not particularly advanced or complicated. The simple square regions we tried may not be optimal at producing this effect—maybe circular ones, or weighted averages would perform better. Furthermore, this technique works to Gaussianize both log-contrast images *and* the original intensity images (data not shown). Fine-tuning of this procedure was not necessary, which suggests that the effects of variance normalization are both basic and robust.

9. Information in the retina

We have seen that natural scenes are statistically quite different from white noise. They are highly structured and correlated over large regions. In the space of all possible images they occupy an infinitesimal volume, a notion which can be quantified using the concept of entropy. The ensemble of natural scenes is quite restricted and thus has much less entropy than a white noise ensemble.

The entropy of natural scenes is an important quantity which is related to how much 'space' is needed to represent them. If every image consisted of a uniform grey, then just one number would specify the whole image: its grey level. On the other hand, if the ensemble really was white noise, then every pixel would need to be retained since they are all totally uncorrelated. Natural images lie somewhere in between.

In order to put a number on the entropy we must invoke an actual image representation which includes noise; otherwise the entropy of a continuous distribution is infinite. For vision the primary representation is at the level of the photoreceptor array, where the scene is first captured. The question of image entropy can be posed as: 'Given the encoding of images by the photoreceptors, what channel capacity is required to transmit them to the brain?' This question is relevant for the visual system since conveying natural images is what the optic nerve does. How much information does each nerve fiber need to be able to send?

Suppose the photoreceptors represent an image as a set of responses $\{y_n\}$. The information in the encoding is

$$I[\{y_n\},\phi] = \int \mathrm{D}\phi\ P[\phi] \int \mathrm{D}\,\{y_n\} P[\{y_n\}|\phi] \log\left[\frac{P[\{y_n\}|\phi]}{P[\{y_n\}]}\right]$$

$$= H[\{y_n\}] - H[\text{noise}] \tag{30}$$

where H is the entropy of a random variable. The second part of this equation is valid under the assumption that the photoreceptor noise is additive and independent of the signal. We use a linear model for the encoding†. First the image is low-pass filtered by the optics according to the diffraction-limited incoherent point-spread function, whose spatial transfer function is (approximately) [12]

$$M(k) = 1 - |k|/k_c \tag{31}$$

for $|k| < k_c$ and zero otherwise. Then the photoreceptors sample this signal, and Gaussian white noise of variance σ^2 is added to their responses. Thus

$$y_n = \int \mathrm{d}^2x\ M(x - x_n)\phi(x) + \eta_n \tag{32}$$

† The responses are more realistically linear in the image intensity, and not the log-contrast. However we find that they both have nearly the same power spectrum, and so the results will be identical. Thus ϕ above may be freely interchanged with I for the purposes of this calculation.

where x_n is the position of the nth receptor, and $\langle \eta_m \eta_n \rangle = \sigma^2 \delta_{m,n}$. This defines the image encoding.

The information per receptor is given by

$$\mathcal{I} = \lim_{N \to \infty} \frac{1}{N} I_N[\{y_n\}, \phi] . \tag{33}$$

An upper bound to this quantity can be found by assuming that the images, and thus the responses, have greater entropy than they actually possess. This is achieved by assigning a Gaussian distribution of images with the same power spectrum measured for natural scenes. This distribution has the greatest entropy consistent with that power spectrum, and so the information is overestimated.

The receptors are placed in a hexagonal arrangement (i.e. on a triangular lattice), as is present in the fovea [25]; they are spaced as far apart as possible so that there is no aliasing, which is nearly the situation in the fovea† [18]. The spacing is thus

$$a = \frac{2}{\sqrt{3}} \frac{\pi}{k_c} . \tag{34}$$

If the receptors were on a square lattice the no aliasing condition would require a 13% increase in the density of receptors, as the spacing would be exactly π / k_c. The area of a unit cell in the hexagonally arranged photoreceptor lattice is $A_c = (\sqrt{3}/2) a^2$.

In the limit of an infinite lattice the Fourier components of a stationary Gaussian signal are independent, and the total information is the sum of the information in each component:

$$\mathcal{I} = \frac{A_c}{4\pi} \int_0^{k_c} \mathrm{d}k \;\; k \log \left[1 + \frac{1}{A_c \sigma^2} |M(k)|^2 S(k) \right] .$$

Here $\mathcal{I}$ is the information per receptor, A_c is the area of the unit cell in the lattice, and σ^2 is the variance of the noise.

We use $S(k) \propto 1/k^{2-\eta}$, with η taking its measured value, and we express the noise level in terms of the SNR in a receptor. It is important to have an information measure to compare with. Ignoring correlations between receptors gives independent Gaussian signals in each receptor, which maximizes the information rate at a given SNR. Spatial redundancy is measured as the difference $\mathcal{I}_{\text{ind}} - \mathcal{I}$, which tells us how much information capacity is effectively wasted due to spatial correlations. For a Gaussian channel at a given SNR (ratio of signal variance to noise variance),

$$\mathcal{I}_{\text{ind}} = \tfrac{1}{2} \log [1 + \text{SNR}] . \tag{35}$$

The quantities are compared in figure 21 for SNR ranging from 1 to 1000. The spatial redundancy is always greater than a factor of two. Perhaps more interestingly, each receptor only conveys a few bits of information per image, which seems quite small. In the fovea there is approximately one ganglion cell fiber leaving the eye for each receptor [24]. If there are 20 new images per second presented to the eye (an upper bound), then each ganglion cell would require an information capacity of about 50 baud, which is well within the 300 baud or so capacity estimated in some neurons [69]. If the one million or so optic nerve fibers must each convey this much information, then the whole optic nerve should operate at about 50 Mbaud, which is five times the capacity of an Ethernet cable. Of course, new random images do not appear every 50ms; instead they are highly correlated over time. So the true figure is certainly much less than this, possibly by many orders of magnitude.

† Such structure is not evident outside the central 1° or so of the human retina, where the situation is complicated by the presence of rods. Cones become much more sparse and randomly arranged in the periphery. Our model receptor lattice has the properties of a very large fovea.

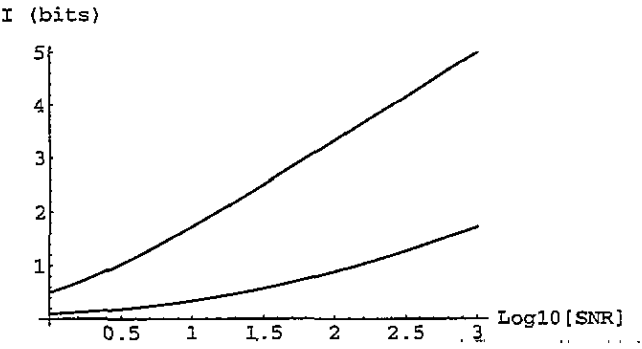

Figure 21. Upper bound to the information per image per receptor as a function of SNR for natural scenes with $\eta = 0.19$ (lower curve), and the information capacity of each receptor (upper curve).

10. Conclusions

Previous work has suggested that natural scenes possess scale invariance. We have also shown this to be the case through the scaling of the power spectrum, local histograms, and pixel information. Our data also demonstrate an approximate invariance to a new symmetry: variance normalization. Images of local pixel variances seem to have the same statistics as the image intensities themselves. Furthermore this process may work recursively, as implied by our data.

Understanding the appearance of scale invariance is straightforward. It seems reasonable that objects in the image can occur at any distance, and some objects even range in size—trees and rocks for instance. When captured in images this means they span many angular scales, and this can produce a scale-invariant ensemble. Another possibility is that the objects and scenes themselves are self-similar, as suggested by the success of fractal image generation [66] and fractal image compression methods [40].

The reasons for the variance normalization invariance are not as intuitive, however. We sought a method for removing the non-Gaussian image statistics. At the same time an invariance appeared. Either we were lucky or there is something fundamental about this type of procedure. It may be a very salient property of natural images that they can be 'Gaussianized' via this iterative scheme. As the nonlinearities present in the visual system begin to be systematically studied, it is important to understand what effect they have on the encoding of images. The outcome of variance normalization suggests that these nonlinearities may prove to be a very rich area of research.

The statistics we have measured are low-dimensional projections of a high-dimensional probability distribution. Consider the beer foam analogy once again. Suppose it has smaller bubbles in the middle than elsewhere, and so the fluid density is greater there. A two-dimensional projection of the foam (i.e. a marginal distribution created by averaging over one variable) will show this increase in density but not the fact that it is made of bubbles. Similarly, one may have to search in very high dimensions in order to visualize the true complexity of the natural image distribution. Finding invariances in the distribution helps us to reduce its intrinsic dimensionality.

While the image statistics are interesting on their own, they also have practical use. Image compression algorithms work through a combination of knowledge about image statistics and psychophysical thresholds [42]. Image restoration procedures ultimately rely

on prior or 'Bayesian' knowledge of the image ensemble [81]. But perhaps the most exciting application of scene statistics is in understanding sensory processing in vision. In order to answer the question 'How well designed is a creature's visual system?', we need three things: a criterion of merit, a set of design constraints, and natural image statistics. Measuring natural scenes is essential to gaining a truly ecological understanding of vision.

Acknowledgments

I am indebted to my collaborators—William Bialek, Horace Barlow, and Chris Wroe—for inspired interactions. I thank Joseph Atick, Roland Baddeley, David Field, Simon Laughlin, Albert Libchaber, Adar Pelah, and Yoav Tadmor for enlightening discussions. This work was supported in part by the Fannie and John Hertz Foundation, the NEC Research Institute, NSF Grant INT-9301746, and The Physiological Laboratory at Cambridge.

Appendix

All images were taken at eye level (about 1.7 m) from random locations in the park. Due to the presence of deer ticks carrying Lyme disease most images were taken from positions on trails; this may cause a systematic bias in image content. The lens was set to focus at infinity, and all shots were at an aperture of $f/5.6$. Due to the optics' limited depth of field, nearby objects would appear out of focus. If any were present in a scene, the azimuthal angle of the camera was changed just enough to remove the offending object from view. The camera's elevation angle was no more than about $\pm 10^{\circ}$ in any image.

Images were gathered using a Sony Mavica MVC-5500 still video CCD camera equipped with a 9.5–123.5 mm zoom lens. This device writes analogue video frames onto small diskettes which are later read off a playback unit (Sony MVR-6500). The video signal is NTSC format RGB. These three signals (red, green and blue) were digitized to 8 bits (0–255) using a Silicon Graphics VideoLab board. To reduce the effects of analogue playback noise we average the result of 32 frame captures to produce floating-point numbers in the range 0–255. Note that since analogue noise is present in the system before digitization, repeated quantization followed by averaging alows us to get around quantization noise.

We use the CIE luminance as a signal. It is derived from the data using the formula [62]

$$Y = 0.59G + 0.3R + 0.11B\,. \tag{A1}$$

This quantity was calibrated against grey cards of known reflectance to give an intensity signal, $I(Y)$, which is proportional to the illuminant flux through the lens. Since the camera had a limited dynamic range, saturation was a possibility. We discard images that have more than about 1% of the pixels saturating. Those pixels that do saturate are set either to the lowest calibrated luminance value or the highest.

References

[1] Arps R B and Truong T K 1994 Comparison of international standards for lossless still image compression *Proc. IEEE* **82** 889–99
[2] Atick J J 1992 Could information theory provide an ecological theory of sensory processing? *Network* **3** 213–51
[3] Atick J J, Li Z and Redlich A N 1992 Understanding retinal color coding from first principles *Neural Comput.* **4** 559–72
[4] Atick J J and Redlich A N 1992 What does the retina know about natural scenes? *Neural Comput.* **4** 196–210
[5] Atick J J and Redlich N 1990 Towards a theory of early visual processing *Neural Comput.* **2** 308

[6] Attneave F 1954 Some informational aspects of visual perception *Psychol. Rev.* **61** 183–93
[7] Barlow H B 1961 Possible principles underlying the transformation of sensory messages *Sensory Communication* ed W A Rosenblith (Cambridge, MA: MIT Press) p 217
[8] Barlow H B 1982 What causes trichromacy? A theoretical analysis using comb-filtered spectra *Vision Res.* **22** 635–43
[9] Barlow H B and Tolhurst D J, Private communication
[10] Bialek W, Ruderman D L and Zee A 1991 Optimal sampling of natural images: A design principle for the visual system? *Advances in Neural Information Processing Systems 3* ed D Touretzky and J Moody (San Mateo, CA: Morgan Kaufmann) pp 363–9
[11] Blakemore C B and Cooper G 1970 Development of the brain depends on the visual environment *Nature* **228** 477–8
[12] Born M and Wolf E 1989 *Principles of Optics* (Oxford: Pergamon) 6th edn
[13] Brillinger D R 1981 *Time Series: Data Analysis and Theory* (San Francisco: Holden-Day)
[14] Buchsbaum G and Gottschalk A 1983 Trichromacy, opponent colour coding and optimum colour information transmission in the retina *Proc. R. Soc.* B **220** 89–113
[15] Burr D C and Morrone M C 1990 Feature detection in biological and artificial visual sytems *Vision: Coding and Efficiency* ed C Blakemore (Cambridge: Cambridge University Press)
[16] Burton G J and Moorhead I R 1987 Color and spatial structure in natural scenes *Appl. Opt.* **26** 157–70
[17] Campbell F W and Green D G 1965 Optical and retinal factors affecting visual resolution *J. Physiol.* **181** 576–93
[18] Campbell F W and Gubisch R W 1966 Optical quality of the human eye *J. Physiol.* **186** 558–78
[19] Cardy J L 1986 conformal invariance and critical behavior *Statistical Physics* ed H E Stanley (Amsterdam: North-Holland) pp 219–24
[20] Carlson C R 1978 Thresholds for perceived image sharpness *Photographic Sci. Eng.* **22** 69–71
[21] Castaing B, Gagne Y and Hopfinger E J 1990 Velocity probability density functions of high Reynolds number turbulence *Physica* **46D** 177–200
[22] Chittka L and Menzel R 1992 The evolutionary adaptation of flower colours and the insect pollinators' colour vision *J. Comp. Physiol.* A **171** 171–81
[23] Cohen R W, Gorog I and Carlson C R 1975 Image descriptors for displays *Technical Report Cont. No* N00014-74-C-0184, Office of Naval Research
[24] Curcio C A and Allen K A 1990 Topography of ganglion cells in human retina *J. Compar. Neurol.* **300** 5–25
[25] Curcio C A and Sloan K R 1992 Packing geometry of human cone photoreceptors: Variation with eccentricity and evidence for local anisotropy *Visual Neurosci.* **9** 169–80
[26] Dannemiller J L 1992 Spectral reflectance of natural objects—how many basis functions are necessary? *J. Opt. Soc. Am.* A **9** 507–15
[27] Daugman J G 1988 Complete discrete 2D Gabor transforms by neural networks for image analysis and compression *IEEE Trans. Acoust., Speech, Signal Process.* **36** 1169–79
[28] Deriugin N G 1956 The power spectrum and the correlation function of the television signal *Telecommunications* **1** (7) 1–12
[29] Dowling J E 1987 *The Retina: An Approachable Part of the Brain* (Cambridge, MA: Belknap Press of Harvard University)
[30] Ebeling W and Pöschel T 1994 Entropy and long-range correlations in literary English *Europhys. Lett.* **26** 241–46
[31] Field D J 1987 Relations between the statistics of natural images and the response properties of cortical cells *J. Opt. Soc. Am.* **4** 2379
[32] Field D J 1994 What is the goal of sensory coding? *Neural Comput.* **6** 559–601
[33] Geman S and Geman D 1984 Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images *IEEE Trans. Patt. Anal. Mach. Intel.* **PAMI-6** 721–41
[34] Hancock P J B, Baddeley R J and Smith L S 1992 The principal components of natural images *Network* **3** 61–70
[35] Heeger D J 1992 Normalization of cell responses in cat striate cortex *Vis. Neurosci.* **9** 181–97
[36] Hirsch H V B and Spinelli D N 1971 Modification of the distribution of receptive field orientation in cats by selective visual exposure during development *Exp. Brain Res.* **13** 509–27
[37] Hsü K J and Hsü A J 1990 Fractal geometry of music *Proc. Natl Acad. Sci. USA* **87** 938–41
[38] Hughes A 1977 The topography of vision in mammals of contrasting life style: comparative optics and retinal organisation *Handbook of Sensory Physiology* vol VII/5, ed F Crescitelli (Berlin: Springer) pp 613–756
[39] Itzykson C and Zuber J-B 1980 *Quantum Field Theory* (New York: McGraw-Hill)
[40] Jacquin A E 1993 Fractal image coding: a review *Proc. IEEE* **81** 1451–465

[41] Jain A K 1989 *Fundamentals of Digital Image Processing* (New York: Prentice Hall) pp 480–83
[42] Jayant N, Johnston J and Safranek R 1993 Signal compression based on models of human perception *Proc. IEEE* **81** 1385–422
[43] Jeng F-C and Woods J W 1991 Compound Gauss–Markov random fields for image estimation *IEEE Trans. Signal Process.* **38** 683–97
[44] Kadanoff L P 1966 Scaling laws for ising models near T_c *Physics* **2** 263–72
[45] Kretzmer E R 1952 Statistics of television signals *Bell Syst. Tech. J.* **31** 751–63
[46] Krinov E L 1947 Spectral reflectance properties of natural formations *Technical Report Technical Translation* TT–439, National Research Council of Canada
[47] Land M F 1990 The design of compound eyes *Vision: Coding and Efficiency* ed C Blakemore (Cambridge: Cambridge University Press) pp 55–64
[48] Laughlin S B 1981 A simple coding procedure enhances a neuron's information capacity Z. *Naturforsch.* **36c** 910–2
[49] Laughlin S B 1992 Retinal information capacity and the function of the pupil *Ophthalm. Physiol. Opt.* **12** 161–4
[50] Laughlin S B 1994 Matching coding, circuits, cells, and molecules to signals—general principles of retinal design in the fly's eye *Prog. Retinal Eye Res.* **13** 165–96
[51] Lenz R 1990 Group invariant pattern recognition *Patt. Recogn.* **23** 199–217
[52] Lenz R 1994 Statistical properties of TV-signals and approximations of the Karhunen–Loéve-transform *Technical Report* LiTH-ISY-R-1617, Department of Electrical Engineering, Linköping University, Sweden
[53] Li Z and Atick J J 1994 Efficient stereo coding in the multiscale representation *Network* **5** 157–74
[54] Li Z and Atick J J 1994 Toward a theory of the striate cortex *Neural Comput.* **6** 127–46
[55] Linsker R 1989 How to generate ordered maps by maximizing the mutual information between input and output signals *Neural Comput.* **1** 402–11
[56] Linsker R 1990 Perceptual neural organization: Some approaches based on network models and information theory *Ann. Rev. Neurosci.* **13** 257–81
[57] Linsker R 1994 Sensory processing and information theory *From Statistical Physics to Statistical Inference and Back* ed P Grassberger and J-P Nadal (Dordrecht: Kluwer) pp 237–247
[58] Lythgoe J N 1979 *The Ecology of Vision* (Oxford: Oxford University Press)
[59] Maloney L T 1986 Evaluation of linear models of surface spectral reflectance with small numbers of parameters *J. Opt. Soc. Am.* A **3** 1673–83
[60] Mollon J D and Bowmaker J K 1992 The spatial arrangement of cones in the primate fovea *Nature* **360** 677–9
[61] Movshon J A and Van Sluyters R C 1981 Visual neural development *Ann. Rev. Psychol.* **32** 477–522
[62] Netravali A N 1988 *Digital Pictures: Representation and Compression* (New York: Plenum)
[63] Ohzawa I, Sclar G and Freeman R D 1985 Contrast gain control in the cat's visual system *J. Neurophysiol.* **54** 651–67
[64] Osorio D and Bossomaier T R 1992 Human cone-pigment spectral sensitivities and the reflectances of natural surfaces *Biol. Cybern.* **67** 217–22
[65] Papoulis A 1991 *Probability, Random Variables, and Stochastic Processes* (New York: McGraw-Hill) 3rd edn
[66] Peitgen H and Saupe D (ed) 1988 *The Science of Fractal Images* (Berlin: Springer)
[67] Procaccia I 1984 Fractal structures in turbulence *J. Stat. Phys.* **36** 649–64
[68] Rieke F, Owen W G and Bialek W 1991 Optimal filtering in the salamander retina *Advances in Neural Information Processing Systems 3* ed D Touretzky and J Moody (San Mateo, CA: Morgan Kaufmann) pp 377–83
[69] Rieke F M, Warland D and Bialek W 1993 Coding efficiency and information rates in sensory neurons *Europhys. Lett.* **22** 151–56
[70] Ruderman D L 1993 Natural ensembles and sensory signal processing *PhD Thesis* University of California, Berkeley
[71] Ruderman D L 1994 Designing receptive fields for highest fidelity *Network* **5** 147–55
[72] Ruderman D L and Bialek W 1992 Seeing beyond the Nyquist limit *Neural Comput.* **4** 682–90
[73] Ruderman D L and Bialek W 1994 Statistics of natural images: Scaling in the woods *Advances in Neural Information Processing Systems 6* ed J D Cowan, G Tesauro and J Alspector (San Mateo, CA: Morgan Kaufmann)
[74] Ruderman D L and Bialek W 1994 Statistics of natural images: scaling in the woods *Phys. Rev. Lett.* 73
[75] Sanger T D 1989 Optimal unsupervised learning in a single-layer linear feedforward neural network *Neur. Networks* **2** 459–473
[76] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379

[77] Shannon C E 1951 Prediction and entropy of printed English *Bell Syst. Tech. J.* **30** 50–64
[78] Snyder A W 1977 Acuity of compound eyes: Physical limitations and design *J. Comp. Physiol.* **116** 161–82
[79] Snyder A W, Stavenga D G and Laughlin S B 1977 Spatial information capacity of compound eyes *J. Comp. Physiol.* **116** 183–207
[80] Srinivasan M V, Laughlin S B and Dubs A 1982 Predictive coding: A fresh view of inhibition in the retina *Proc. R. Soc.* B **216** 427–59
[81] Stark H (ed) 1987 *Image Recovery: Theory and Application* (New York: Academic)
[82] Tolhurst D J, Tadmor Y and Chao T 1992 Amplitude spectra of natural images *Opthal. Physiol. Opt.* **12** 229–32
[83] van Hateren J H 1992 Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation *J. Comp. Physiol.* A **171** 157–70
[84] van Hateren J H 1992 A theory of maximizing sensory information *Biol. Cybern.* **68** 23–9
[85] Voss R F and Clarke J 1978 $1/f$ noise in music: Music from $1/f$ noise *J. Acoust. Soc. Am.* **63** 258–63
[86] Webber C J S 1991 Competitive learning, natural images and cortical cells *Network* **2** 169–87
[87] Webster R J 1993 Ambient noise statistics *IEEE Trans. Sig. Proc.* **41** 2249–53
[88] Wiener N 1949 *Time Series* (Cambridge, MA: MIT Press)
[89] Wilson K G 1983 The renormalization group and critical phenomena *Rev. Mod. Phys.* **55** 583